Vers l'application de l'apprentissage par renforcement inverse aux réseaux naturels d'attention

Bertille Somon[1,2], Aurélien Fermo[1,3], Frédéric Dehais[2,1], Caroline P. C. Chanel[2,1]

[1] ANITI, Artificial and Natural Intelligence Toulouse Institute, France
[2] ISAE-SUPAERO, Université de Toulouse, France
[3] ENS-PSL, Département d'Études Cognitives, Paris, France

Résumé

Le cerveau humain, pour allouer de manière optimale les ressources attentionnelles limitées dont il dispose, supprime ou renforce l'activation de circuits neuronaux : il implémente des heuristiques. Dans une approche novatrice, nous proposons d'utiliser l'apprentissage par renforcement inverse pour caractériser la dynamique d'activation de ces réseaux. Un protocole expérimental est proposé, et les données collectées devraient permettre, à terme, de vérifier cette démarche.

Mots-clés

Processus Décisionnels de Markov, Apprentissage par renforcement inverse, Electroencephalographie, Connectivité dirigée

Abstract

The human brain possesses limited attentional resources and requires heuristics in order to optimise their allocation through neural networks reinforcement. We propose here that using inverse reinforcement learning is an interesting approach to characterize the dynamic activation of these networks. We present an experimental setting aiming at this characterization, and we propose that data collection will comfort this position.

Keywords

Markov Decision Process, Inverse Reinforcement Learning, Electroencephalography, Directed Connectivity

1 Introduction

La répartition optimale de l'attention est une question essentielle dans nos activités multi-tâches de la vie quotidienne. Elle s'appuie sur un compromis entre des politiques d'exploration et d'exploitation des flux d'information pertinents, c'est-à-dire qu'elle consiste à focaliser et maintenir l'attention tout en la laissant permissive aux changements inattendus [15, 9]. À la lumière de certaines études, les corrélats neuronaux de cette dynamique attentionnelle ont pu, en partie, être identifiés. Notamment, les mécanismes attentionnels descendants et ascendants (dits *top-down* et *bottom-up*) sont respectivement délimités par des réseaux cérébraux dorsaux et ventraux. Ces réseaux sont eux-mêmes en étroite interaction avec le cortex cingulaire antérieur responsable de l'allocation des ressources [15, 9]. En condition normale, l'attention est dite "divisée" et permet de traiter efficacement les informations de l'environnement dans différentes modalités (e.g. visuelle, auditive, tactile). Aussi des mécanismes neuronaux oscillatoires sont-ils mis en œuvre pour synchroniser et augmenter l'activité des ces réseaux qui traitent les informations les plus importantes. Enfin, des mécanismes de phasage permettent d'alterner et de rythmer le traitement des informations liées à des tâches secondaires ou inattendues [13].

Toutefois, il a été avancé que la fatigue, un stress intense ou une importante charge de travail pouvaient entraîner un déficit de l'homéostasie entre ces réseaux attentionnels et conduire à une attention dite focalisée, voire, dans les cas extrêmes, "tunnélisée" (pour un revue voir [10]). Concrètement, ces situations dégradées entraînent surtout la suppression de l'activité liée aux tâches secondaires et *bottom-up*, laquelle joue pourtant le rôle primordial d'alerter le cerveau en cas d'imprévus [14, 33, 10]. Bien que ce mécanisme permette, à la manière d'un fusible, de prévenir la surcharge mentale et d'éviter la distraction de l'attention dans des situations complexes, l'omission d'informations essentielles peut avoir des conséquences dévastatrices dans des scénarios de la vie réelle. Ce phénomène, appelé cécité ou surdité attentionnelle, est reconnu pour être à l'origine d'accidents de la route et dans l'aéronautique lorsque des stimuli visuels (e.g. un autre véhicule) ou auditifs (e.g. une alarme) ont pu être négligés [11].

La compréhension et la caractérisation de ces dynamiques cérébrales représentent donc un enjeu de recherche important. En pratique, le suivi en ligne des connectivités cérébrales pourrait permettre le développement d'interfaces cerveau-machine capables de détecter, en situation opérationnelle, des états attentionnels dégradés [12]. Ensuite, l'étude des mécanismes cérébraux et des heuristiques mises en œuvre par la biologie de l'évolution pourraient en retour inspirer de nouveaux algorithmes d'intelligence artificielle [24]. Dans cette perspective, nous pensons qu'il serait pertinent d'appliquer le cadre formel de l'apprentissage par renforcement inverse (IRL [27]) à l'identification des corrélats neuronaux de la dynamique attentionnelle. En particulier, nous suggérons de nous appuyer sur la notion de *fonction de récompense* pour modéliser les stratégies d'activa-

tion ou de suppression de réseaux attentionnels que le cerveau est susceptible de mettre en oeuvre. Dès lors serions-nous en mesure de mieux caractériser les *politiques* attentionnelles d'un ensemble d'agents et de distinguer formellement celles qui sont efficaces (au sein d'une population dite experte) de celles qui ne le sont pas (population dite novice).

1.1 Travaux antérieurs

De nombreuses études se sont attachées à comprendre les liens de causalité ou de corrélation qui peuvent exister entre les différentes aires cérébrales, améliorant ou perturbant les changements d'attention sélective inter-modale [1] [5]. De plus en plus d'évidences montrent que les influences inter-modales sur les cortex sensoriels primaires, responsables de la détection de stimuli (par exemple auditifs ou visuels), sont modulées par la synchronisation d'oscillations neuronales grâce à une ré-initialisation des phases ou bien un entraînement neuronal (i.e. le processus au travers duquel deux ou plusieurs oscillateurs auto-entretenus sont couplés et se synchronisent), ou encore une combinaison de ces deux mécanismes [5]. Les mesures de connectivité cérébrale permettent de mettre en évidence ces communications (dirigées ou non) entre plusieurs aires.

Dans le domaine des neurosciences, la connectivité dite *effective* permet d'identifier les réseaux fonctionnels qui varient dans le temps par des techniques basées entre autres sur la causalité de Granger. Plusieurs études ont utilisé ces métriques de connectivité effective comme entrée d'algorithmes d'apprentissage (voir par exemple [19] pour une revue) permettant de transférer et tester ces résultats en ligne dans différentes conditions. Ces algorithmes permettent d'obtenir des résultats de classification de différents états cognitifs avec des performances très élevées [12].

Par ailleurs, les mesures de connectivité permettent d'établir des graphes de connectivité dirigés et dynamiques. Les graphes sont alors définis par un ensemble de noeuds et de connections qui permettent d'obtenir une représentation abstraite des éléments d'un système et de leurs interactions [7]. Appliquée à l'électroencéphalographie (EEG), la théorie des graphes permet de définir des liens entre l'activité cérébrale de différentes aires, par seuillage ou validation statistique des mesures de connectivité. Il est à noter que des couples de connectivité seuillés peuvent définir une matrice d'adjacence de graphe. Il est alors possible d'obtenir des noeuds et des arrêtes représentant les aires cérébrales significativement actives et les métriques de connectivité qui les relient de manière significative. Cette approche permet d'obtenir une modélisation dynamique en considérant que chaque réseau fonctionnel représente un état d'un système dynamique [18, 7]. Plus récemment, cette approche à été étendue à l'utilisation de Chaînes de Markov pour modéliser la dynamique de ce système [34].

Nous posons que la dynamique d'états (réseaux fonctionnels d'attention) est régie par une politique optimisant l'allocation des ressources. Approximer cette dynamique par des chaînes de Markov dites contrôlées serait une approche envisageable. Toutefois, comme le soulignent Ng et Russell [27], l'apprentissage par renforcement (RL) reste peu applicable à des cas concrets de simulation du comportement humain puisque la fonction de récompense y est supposée connue. C'est pourquoi l'apprentissage par renforcement *inverse* (IRL) nous paraît être le cadre le plus approprié si l'on veut expliciter la dynamique et la qualification d'action (fonction de récompense et politique associée).

1.2 Problématique

Dans ce contexte, l'objectif de l'étude que nous souhaitons mener est de caractériser l'efficacité de la dynamique cérébrale de l'attention sélective par le biais de mesures d'activité EEG lors d'une tâche expérimentale contrôlée de changement attentionnel inter-modal. Plus précisément, nous supposons la dynamique attentionnelle guidée par une politique cérébrale intrinsèque et nous pensons pouvoir la décrire formellement en nous fondant sur des données EEG et comportementales collectées. L'apprentissage de la fonction de récompense (générant cette politique) par IRL permettra de comprendre la sélection des actions. Étant capable, grâce à cette fonction de récompense, de représenter le degré d'efficacité d'une politique attentionnelle menée par un agent, nous pourrons ainsi mieux prédire les performances cognitives et aider au développement d'algorithmes d'intelligence artificielle qui soient bio-inspirés [24].

2 IRL et dynamique cérébrale

L'IRL appliqué à un Processus Décisionnel de Markov (MDP) a été caractérisé de manière informelle par Russel [27] comme la définition de la fonction de récompense qu'il est nécessaire d'optimiser connaissant : i) des mesures du comportement d'un agent au cours du temps et dans différentes circonstances, ii) si nécessaire, des mesures des entrées sensorielles de cet agent, et iii) si possible, un modèle de l'environnement [27].

Rappelons d'abord la définition formelle d'un MDP. Un MDP à horizon infini est un n-uplet $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$, où $\mathcal{S}$ est l'ensemble d'états ; $\mathcal{A}$ est l'ensemble d'actions ; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ est la fonction de transition d'état qui spécifie la probabilité de transiter vers l'état $s' \in \mathcal{S}$ depuis l'état $s \in \mathcal{S}$ quand l'action $a \in \mathcal{A}$ est réalisée, telle que $\mathcal{T}(s', a, s) = p(s'|s, a)$; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ est la fonction de récompense qui définit une récompense $r(s, a)$ quand l'action $a \in \mathcal{A}$ est prise dans l'état $s \in \mathcal{S}$; enfin, $\gamma \rightarrow [0, 1[$ est le facteur d'oubli. Pour résoudre un MDP à horizon infini, il suffit de rechercher une politique markovienne déterministe stationnaire $\pi : \mathcal{S} \rightarrow \mathcal{A}$ qui maximise la fonction de valeur généralement définie comme :

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \mid s_0 = s\right] \qquad (1)$$

1. L'attention sélective inter-modale correspond à l'attention portée de manière sélective et alternée sur différentes modalités sensorielles lorsqu'elles sont présentées simultanément.

Cette équation peut être développée et ré-écrite en tant que :

$$\begin{aligned} V^{\pi}(s_0) &= \mathbb{E}\left[\gamma^0 r(s_0, \pi(s_0))\right] + \\ &\quad \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t)) \mid s_t = s_1, s_0, \pi(s_0)\right] \\ V^{\pi}(s_0) &= r(s_0, \pi(s_0)) + \gamma \sum_{s \in S} p(s|s_0, \pi(s_0)) V^{\pi}(s) \end{aligned}$$

Ceci indique que la politique optimale (stationnaire déterministe markovienne) π^* peut être calculée, ainsi que la fonction de valeur optimale $V^* = V^{\pi^*}$, sur la base de l'équation de Bellman :

$$V^*(s) = \max_{a \in \mathcal{A}} \left[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^*(s') \right] \quad (2)$$

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \left[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^*(s') \right] \quad (3)$$

Il est à noter que la valeur Q de l'état s et de l'action a sous la fonction de valeur V^{π^*}, dénotée $Q^{\pi^*}(s,a)$, telle que :

$$Q^{\pi^*}(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^{\pi^*}(s') \quad (4)$$

est la valeur anticipée d'une étape de l'action a dans l'état s en suivant la politique optimale π^* et en obtenant V^*, la vraie valeur optimale attendue. Nous précisions que la valeur d'un état peut alors être calculée par :

$$V^*(s) = \max_{a \in \mathcal{A}} Q^{\pi^*}(s,a)$$

De nombreux algorithmes de solution exacte ou approchée de MDP ont été proposés dans la littérature (voir [21]). Les algorithmes les plus récents explorent des heuristiques pour guider la recherche dans l'espace d'état vers des régions qui sont plus susceptibles d'être visitées par la politique optimale. Ainsi la valeur d'un état, ou la Q-valeur d'une action dans un état, peut être approchée de façon efficace.

L'apprentissage par renforcement (RL) est un des moyens de résolution des MDP où le modèle descriptif n'est pas connu. À la place, il est admis qu'un modèle génératif ou un simulateur de l'environnement est disponible. Le but de l'agent MDP est alors d'approcher la fonction de valeur (et par conséquent la politique) en interagissant avec l'environnement par exploration ou exploitation (ou les deux) des séquences d'actions, tout en évaluant les récompenses obtenues en moyenne. La valeur d'une action (Q-valeur) est alors approchée par diverses méthodes (voir [31]) fondées sur des modèles estimés ou non.

L'IRL, quant à lui, s'intéresse à l'apprentissage de la fonction de récompense [27]. Cela est généralement réalisé sur la base des trajectoires (état-action) effectuées par un agent (en MDP), l'objectif étant de comprendre quelle fonction de récompense a guidé la politique de cet agent dont on observe le comportement.

Initialement, le développement de l'IRL a été motivé par le fait que les recherches sur le RL étaient généralement peu applicables à des cas concrets, et d'autant moins au comportement humain, puisqu'elles avaient tendance à supposer que les fonctions de récompense étaient fixes et connues [27]. Or, dans la majorité des cas, les fonctions de récompense permettant à un agent d'agir avec succès dans son environnement ne sont pas prédéfinies. Par ailleurs, elles sont aussi plus complexes que les fonctions de récompense généralement utilisées en RL et varient d'un agent à un autre. C'est pourquoi il est nécessaire de les inférer selon l'environnement d'application mais aussi selon le type d'agent visé. L'un des intérêts majeurs est que les fonctions de récompense sont notamment plus robustes, rapides et transférables que la politique de l'agent telle que définie dans le contexte du RL [27, 2].

Ainsi l'IRL est-il proposé comme une façon de résoudre un problème d'apprentissage lorsque la fonction de récompense n'est pas supposée connue. L'une des premières motivations en faveur de l'IRL vient de l'intérêt pour l'inférence des objectifs, stratégies ou intentions qui président au comportement animal [27]. Dans le cadre du RL classique, lorsque les intentions d'un agent ne sont pas connues, modéliser son comportement nécessite, après avoir entraîné le modèle, de re-paramétrer la fonction de récompense et d'entraîner à nouveau le modèle jusqu'à obtenir le comportement désiré. Au contraire, l'IRL entend inférer automatiquement cette fonction de récompense en prenant en entrée un ensemble d'observations émanant d'un expert (ou simplement de l'agent dont on veut connaître les objectifs).

Le cadre de résolution de l'IRL est un MDP $\mathcal{M}\backslash_{\mathcal{R}}$: un modèle MDP classique mais dont on a retiré la fonction de récompense. L'ensemble des observations qui proviennent de l'agent dont on veut connaître la fonction de récompense est dénoté $\mathcal{D} = \{\tau_0, \tau_1, \tau_2, ...\tau_n\}$, $n \in \mathbb{N}$ avec $\tau_i = \{(s_0, a_0), (s_1, a_1), ..., (s_k, a_k)\}_i$, $s_k \in \mathcal{S}$, $a_k \in \mathcal{A}$, $i \in \mathbb{N}$ une trajectoire. L'IRL prend donc en entrée le couple $\{\mathcal{M}\backslash_{\mathcal{R}}, \mathcal{D}\}$ et donne en sortie $\hat{R}^*$, la fonction de récompense estimée de l'agent. Bien que retrouver la fonction de récompense ne soit pas nécessairement l'objectif principal de tous les algorithmes d'IRL, il apparaît que la grande majorité des algorithmes suppose que $R^*(s, \pi(s)) \in \mathcal{H}_\phi(s, \pi(s))$, où $\mathcal{H}_\phi(s, \pi(s)) = \{\boldsymbol{\theta}^T \boldsymbol{\phi}(s, \pi(s)), \boldsymbol{\theta} \in \mathbb{R}^n\}$ est l'espace d'hypothèses de la fonction de récompense. En d'autres termes la plupart des algorithmes d'IRL supposent que chaque fonction de récompense que l'on veut tester est représentée par une combinaison linéaire des n propriétés (*features*) d'une paire état-action :

$$\begin{aligned} r(s, \pi(s)) &= \theta_0 \phi_0(s, \pi(s)) + \ldots + \theta_n \phi_n(s, \pi(s)) \\ &= \boldsymbol{\theta}^T \boldsymbol{\phi}(s, \pi(s)) \end{aligned}$$

Ainsi, nous cherchons à représenter la performance d'une politique comme suit :

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{\theta}^T \boldsymbol{\phi}(s_t, \pi(s_t)) \mid s_0 = s\right] \\
&= \boldsymbol{\theta}^T \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{\phi}(s_t, \pi(s_t)) \mid s_0 = s\right] \\
&= \boldsymbol{\theta}^T \boldsymbol{\mu}^{\pi}(s)
\end{aligned}
$$

Pour une fonction de récompense donnée, $\boldsymbol{\mu}^{\pi}(s)$ est l'espérance de propriétés (*feature expectations*) obtenue en suivant la politique π associée. Ainsi l'objectif commun aux algorithmes d'IRL est de trouver le vecteur de paramètres optimal $\boldsymbol{\theta}^*$ tel que : $\boldsymbol{\theta}^{*T}\boldsymbol{\mu}^{\pi^*}(s) \geq \boldsymbol{\theta}^{*T}\boldsymbol{\mu}^{\pi}(s), \forall \pi \in \Pi$. Il faut trouver $\boldsymbol{\theta}^*$ tel que la performance (définie ci-dessus) de la politique qui lui est associée soit supérieure à celle de n'importe quelle autre politique.
D'où si deux politiques partagent les mêmes espérances de propriétés alors leurs valeurs sont identiques [1] : $\boldsymbol{\mu}_1^{\pi}(s) = \boldsymbol{\mu}_2^{\pi}(s) \Rightarrow \boldsymbol{\theta}^T\boldsymbol{\mu}_1^{\pi} = \boldsymbol{\theta}^T\boldsymbol{\mu}_2^{\pi} \Rightarrow V^{\pi_1}(s) = V^{\pi_2}(s)$. C'est pourquoi la plupart des algorithmes d'IRL cherchent à minimiser une fonction de perte $\mathcal{L}$ tel que :

$$
\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\mu}^{\pi^*}, \boldsymbol{\mu}^{\boldsymbol{\pi_\theta}}) \tag{5}
$$

L'objectif est donc de réduire, par itérations successives, la distance entre les espérances de propriétés de l'agent (appelé *expert*) dont on veut inférer la fonction de récompense et celles qu'on trouve à chaque itération de l'algorithme.
Soulignons que la minimisation de la fonction $\mathcal{L}$ peut se faire de différentes manières : (i) le *feature matching* dont l'objectif est précisément de faire correspondre les espérances de propriétés de l'algorithme avec celles de l'expert en réactualisant les valeurs de $\boldsymbol{\theta}$ à chaque itération [1, 32, 35] ; (ii) le *max-margin* où l'on optimise sous contrainte l'espérance des récompenses des trajectoires de l'expert observées tel qu'elle soit supérieure à celle qui est obtenue en suivant n'importe quelle autre politique [29, 22, 20] ; (iii) en représentant la fonction de récompense comme le simple paramètre conditionnant en probabilité une classe de politiques et en maximisant cette probabilité par inférence bayésienne, méthode de gradient, etc. [28, 26, 4]
D'une manière générale, les algorithmes d'IRL suivent les étapes suivantes afin d'inférer la fonction de récompense :

1. définition de l'ensemble des observations $\mathcal{D}$ et du modèle $\mathcal{M}_{\mathcal{R}}$;
2. initialisation des paramètres de la fonction de récompense ;
3. résolution de $\mathcal{M}$ par RL classique sous l'hypothèse de la fonction de récompense courante ;
4. calcul de la fonction de perte $\mathcal{L}(\boldsymbol{\mu}^{\pi^*}, \boldsymbol{\mu}^{\boldsymbol{\pi_\theta}})$ puis optimisation des paramètres de la fonction de récompense pour minimiser la fonction de perte ;
5. répéter 3. et 4. jusqu'à réduire la divergence en dessous d'un seuil fixé [2].

Notre but est, en nous appuyant sur une expérimentation assez fondamentale (présentée dans la section suivante), d'approcher au mieux les fonctions de récompenses des participants, jugés performants ou non selon leurs réponses, pour mieux expliquer la politique intrinsèque de changement (inter-modal auditif/visuel) attentionnel mise en œuvre par leur cerveau.

3 Protocole expérimental

L'objectif de cette étude est de caractériser la dynamique cérébrale associée à l'attention sélective grâce à des mesures d'activité EEG lors d'une tâche de changement attentionnel inter-modal. Nous proposons d'enregistrer l'activité EEG de participants en les soumettant à des stimulations audiovisuelles durant lesquelles ils devront porter leur attention sur la modalité visuelle, la modalité auditive ou bien changer (condition « *switch* ») entre les deux modalités. Une tâche de mémoire de travail (*N-back*[2] [30]) sera incluse dans chaque modalité et permettra de définir (i) le moment où le participant doit changer de modalité sensorielle dans la condition *switch*, et (ii) les performances du participant dans toutes les conditions en termes de précision et de temps de réaction. Les performances des participants dans la condition auditive et dans la condition visuelle permettront de définir deux groupes : un groupe expert – aux performances élevées – et un groupe novice. Des métriques de l'activité cérébrale associées aux transitions inter-modales permettront de caractériser les corrélats neuronaux des changements d'orientation attentionnelle pour chaque modalité (visuelle et auditive) mais aussi de modéliser les transitions efficaces et inefficaces à la fois pour la population experte et pour les novices.
Les métriques les plus pertinentes seront alors utilisées pour la définition des graphes, lesquels constitueront des états. Les actions, ainsi que la fonction de transition d'un modèle MDP, seront alors définies et approchées respectivement en fonction des trajectoires d'état observées. La Figure 1 illustre la modélisation envisagée. La fonction de récompense sera alors estimée par une méthode d'IRL permettant non seulement de caractériser la politique des changements attentionnels visuels et auditifs, mais aussi de discriminer les deux populations. Le protocole expérimental détaillé dans la suite a reçu l'avis favorable du Comité d'Éthique de la Recherche (CER) de l'Université Fédérale Toulouse Midi-Pyrénées (CER $2020 - 322$).

3.1 Matériel et Méthode

3.1.1 Participants

La littérature portant sur l'attention sélective uni- et inter-modale fait référence à des calculs de puissance statistique [23] qui permettent d'établir un minimum de 20 participants pour observer l'effet de la modalité (visuelle, auditive ou audiovisuelle) ainsi que l'effet de l'expertise (experts vs.

2. La tâche de N-back est une tâche de mémoire de travail où l'on présente des stimuli successifs au participant qui doit répondre quand un stimulus a déjà été présenté *N* positions auparavant. Par exemple dans une tâche de 2-back, avec la séquence *M V A V B*, le participant doit répondre au deuxième V qui est présenté deux essais après le premier.

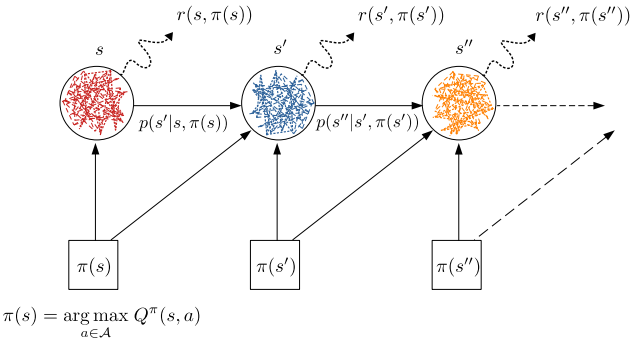

FIGURE 1 – Modèle MDP envisagé représentant la dynamique attentionnelle des participants. Chaque graphe de connectivité, obtenu sur la base des données EEG, constitue un état s. $\pi(s)$ est l'action prise par le participant (la réponse aux stimuli) qui, dans un état s, fait transiter vers un autre état attentionnel s' avec probabilité $p(s'|s,a)$, engendrant la récompense r.

novices). Pour notre propre expérience nous en recruterons 30 au minimum en prévision d'éventuelles pertes de données lors des enregistrements.

3.1.2 Stimuli

Les stimuli visuels consistent en un damier noir et blanc, modulé par une onde sinusoïdale à $48Hz$ (voir fig. 2a), et en une série de points rouges qui apparaissent au centre du damier, à raison de deux, trois ou quatre occurrences successives. Les stimuli auditifs consistent en un son sinusoïdal à $500Hz$, modulé par un son à $40Hz$ (voir fig. 2c), et en de brèves augmentations de 200% de l'intensité de cette modulation, à raison de deux, trois ou quatre occurrences successives. Chaque stimulus (visuel ou auditif) est présenté sur une durée de 2 cycles et permet d'effectuer la tâche de N-back.

3.1.3 Tâche et conditions expérimentales

Dans le but d'obtenir des performances convenables et d'associer l'absence de réponse à une absence de détection des stimuli de changement de modalité (*switch*), une tâche de 0-back est effectuée. Les participants doivent donc détecter les trains de deux stimulations consécutives (i.e. les cibles) et appuyer sur un bouton le cas échéant (une touche spécifique étant associée à chaque type de stimulus).
Les stimuli visuels et auditifs sont présentés de manière simultanée et continue par bloc de 3 minutes. Chaque bloc contient donc environ 40 trains visuels et 40 trains auditifs. Trois conditions expérimentales sont alors déterminées par la présentation de trois types de bloc :

- des blocs durant lesquels le participant doit focaliser son attention sur la tâche visuelle sans tenir compte des stimulations auditives (5 blocs au total) ;
- des blocs auditifs, où la consigne est inversée par rapport aux blocs visuels (5 blocs au total) ;
- des blocs inter-modaux visuo-auditifs, où les stimuli cibles entraînent un changement de focus attentionnel entre le visuel et l'auditif de manière alternée (10 blocs au total).

Pour chaque condition, 30% des trains sont des cibles ($\sim$ 240 cibles au total dont 120 représentent des changements inter-modaux). La figure 2e est une représentation schématique des enchaînements de trains lors des blocs de changement attentionnel.

3.1.4 Recueil et analyse des données

Les données comportementales et EEG sont recueillies de manière continue tout au long des 20 blocs d'expérimentation, et seront prétraitées selon les standards en vigueur.

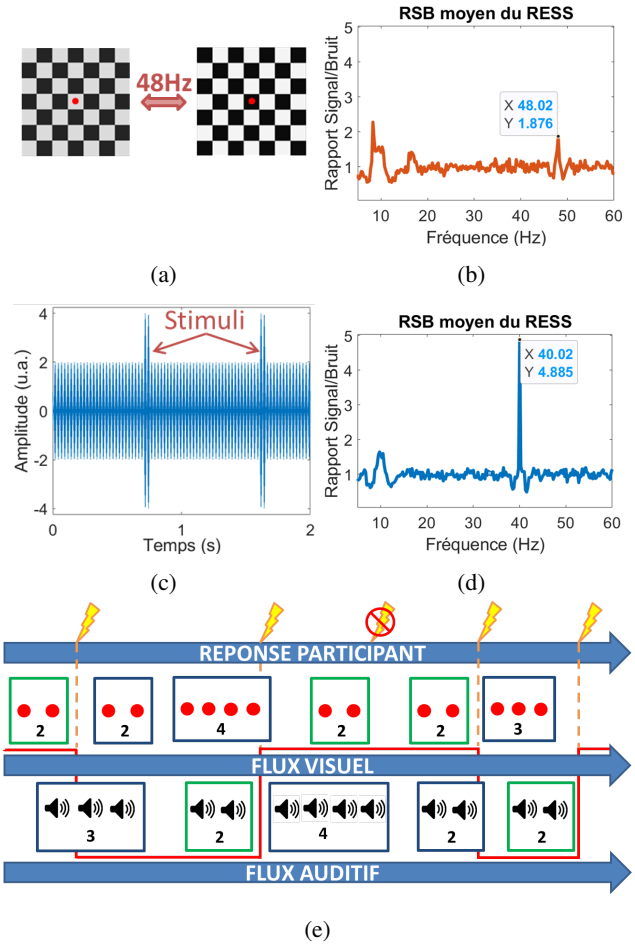

FIGURE 2 – Stimulations (a) présentées à $48Hz$ dans le domaine visuel et (c) modulées à $40Hz$ dans le domaine auditif afin de déclencher des activités cérébrales de *steady-states* (b) visuels à $48Hz$ et (d) auditifs à $40Hz$. (e) Description de la tâche de changement attentionnel inter-modal durant laquelle le flux auditif (dernière ligne) et le flux visuel (ligne intermédiaire) sont présentés simultanément. Lors des blocs *switch* le participant doit appuyer sur une touche du clavier (réponse participant en haut) et changer son attention pour l'autre modalité lorsqu'une cible (deux points ou deux sons – en vert sur le schéma) est présentée. Il doit ensuite rester focalisé sur la même modalité jusqu'à la prochaine cible dans cette modalité. Le trait rouge indique le "chemin attentionnel" du participant en fonction des stimuli et de ses réponses ou absences de réponse (e.g. à la troisième cible).

Comportementales : Les taux d'erreurs et les temps de réaction moyens seront analysés après le déroulement de l'expérience à l'aide d'analyses de variance (ANOVA) à mesures répétées. Les taux d'erreurs permettront de définir les deux groupes (experts et novices).

Electrophysiologiques : Deux types de mesures EEG seront extraites : des mesures de puissance fréquentielle et des mesures de connectivité. Concernant la puissance fréquentielle, la méthode RESS (*Rythmic Entrainment Source Separation* [8]) sera appliquée afin d'identifier les sources cérébrale émettrices des activités de *steady-states*[3] visuels et auditifs habituellement observés dans ce type de tâche, de manière guidée. Les résultats d'une telle mesure observés lors de pré-tests à l'expérimentation sont présentés pour la modalité visuelle (fig. 2b) et pour la modalité auditive (fig. 2d) à titre d'exemples. La quantification dynamique sera obtenue en extrayant la magnitude de la transformée de Hilbert sur la fréquence du *steady-state* au cours du temps. Ces deux mesures seront comparées dans les différentes conditions expérimentales grâce à une ANOVA à mesures répétées.

4 Connectivité et graphes

Dans chaque condition, des métriques de connectivité dirigée seront extraites sur des fenêtres temporelles de 5 secondes, centrées sur la fin de la stimulation cible, afin de rendre compte des activations ou suppressions des réseaux attentionnels lors des passages de la modalité visuelle à l'auditive, et inversement. Pour cela, les modèles d'autorégrésseurs multi-variées (ou Multivariate Autoregressive (MVAR) models) seront utilisés.

Soit $X(t) = \sum_{d=1}^{p} A_{ij}(d)X(t-d) + e(t)$ le modèle MVAR, avec $X(t)$ une série temporelle représentant l'activité cérébrale aux différentes k électrodes, $A_{ij}(d)$ la matrice de taille $k \times k$ des coefficients d'estimation du modèle d'ordre p à chaque instant d associée au couple d'électrodes ij, et $e(t)$ l'erreur de prédiction. Il est possible de calculer la fonction de transfert $\hat{H}(f)$ du modèle MVAR dans le domaine fréquentiel telle que [17] :

$$\hat{H}(f) = \left(\sum_{d=0}^{p} A(d) e^{-2i\pi f \Delta t} \right)^{-1}$$

où l'élément $H_{ij}(f)$ de la matrice $\hat{H}(f)$ décrit la connexion entre la $j^{\text{ème}}$ entrée (électrode) et la $i^{\text{ème}}$ sortie (électrode) du système, et Δt est un intervalle de temps.

La fonction de transfert spectrale ainsi que la matrice de corrélation permettent d'estimer différentes métriques de connectivité : la fonction de transfert dirigée (*Directed Transfer Function* ou DTF) et la Cohérence Partielle Dirigée (*Partial Directed Coherence* ou PDC) ainsi que leurs métriques affiliées (e.g. la *full frequency* DTF). Ces mesures représentent respectivement le flux d'entrée ou le flux de sortie causal de l'électrode j vers i.

3. Les *steady-states* sont définis comme des activités oscillatoires résultant de stimulations répétées et pour lesquelles les neurones se synchronisent à la fréquence de stimulation [16].

De nombreuses mesures de connectivité s'affranchissent de l'aspect temporel, nécessaire à l'estimation de la dynamique attentionnelle, en passant au domaine fréquentiel. Pour pallier cette difficulté, des métriques dépendantes du temps ont été développées : ce sont les métriques dénommées *short-time* [18]. Lorsque plusieurs répétitions d'une même stimulation sont disponibles (un nombre de répétitions $r = 1, 2, \ldots, N_T$), il est possible de définir les mesures de connectivité précédentes en moyennant les matrices de corrélation à travers les différents essais sur de courtes fenêtres (avec N_S points temporels) considérées comme quasi-stationnaires [18]. On obtient alors une matrice de corrélation croisée dépendante du temps ($\widetilde{R}_{ij}(s)$ avec s un délai pré-défini) telle que :

$$\widetilde{R}_{ij}(s) = \frac{1}{N_T} \sum_{r=1}^{N_T} \frac{1}{N_S} \sum_{t=1}^{N_S} X_i^{(r)}(t) X_j^{(r)}(t+s)$$

cette matrice $\widetilde{R}_{ij}(s)$ permet alors d'obtenir une DTF à court terme (Short-time DTF ou SDTF) par exemple.

Dans tous les cas (mesures statiques ou dynamiques), les matrices d'adjacence sont calculées sur les métriques de connectivité les plus appropriées. Ici, chaque graphe défini sur un temps court ou non représente un état défini à l'échelle du participant. Puis, pour une action donnée, les graphes sont comparés à l'échelle du groupe (inter-participants). Les graphes (i.e. les états) similaires au travers des participants pourront alors être agrégés soit par une méthode basée sur une approche de mélange assortatif adaptée des statistiques de graphes pour les données EEG [18], soit par une méthode de clustering [7]. Enfin, les divergences observées à l'échelle du groupe entre les états inter-participants pour chaque condition expérimentale seront capturées par la fonction de transition du modèle MDP.

4.1 Application de l'IRL

La construction des graphes, obtenus sur des courtes fenêtres temporelles, nous permettra de définir les états du MDP sur lequel nous voulons appliquer les algorithmes d'IRL. Les métriques de connectivité de chaque graphe constitueront le vecteur $\boldsymbol{\theta}$ de propriétés (*features*) définissant chaque état, et les réponses des participants aux stimuli constitueront les actions dans le MDP (voir fig. 1). *In fine* nous obtiendrons pour chaque participant l'ensemble de trajectoires $\mathcal{D}$ dont nous avons besoin pour inférer la fonction de récompense $\mathcal{R}$. Le design expérimental présenté ainsi que l'objectif final contraignent le choix des algorithmes d'IRL qu'il est possible d'utiliser. Ni les algorithmes d'*apprenticeship* [1, 32, 29], ni ceux qui envisagent la sous-optimalité du comportement de l'expert [26, 6] ne semblent appropriés puisqu'ici nous souhaitons inférer la fonction de récompense elle-même d'une population, et non purement copier une politique ou encore extrapoler autour d'elle.

Nous supposons que : i) nos espaces d'états et d'actions sont discrets et nos vecteurs $\boldsymbol{\theta}$ de propriétés pour chaque état bien définis par des mesures observées ; ii) il est plus pertinent d'inférer une distribution de probabilité sur l'en-

semble des fonctions de récompense qui peuvent expliquer la dynamique cérébrale et comportementale d'un participant ; et iii) l'expérience permet de fournir en entrée d'un algorithme des trajectoires provenant non pas d'un seul mais d'une trentaine d'agents différents aux performances variées.

Par conséquent, nous portons actuellement notre intérêt sur l'algorithme de Babeş-Vroman [3] qui est l'un des seuls à utiliser à la fois un modèle probabiliste (bayésien) et une méthode de *clustering* (*Expectation-Maximization*) capable de définir plusieurs groupes d'agents selon la distribution de probabilité sur les fonctions de récompense qui leur est associée. En entrée l'algorithme prend l'ensemble des trajectoires $\mathcal{D}$ dont on sait qu'elles proviennent d'intentions différentes (ici de stratégies attentionnelles) et le nombre m de *clusters* maximal supposé (avec $|\mathcal{D}| = n > m$). On initialise le vecteur $\mathbf{\Theta} = (\rho_1, ..., \rho_m, \theta_1, ..., \theta_m)$, où ρ_j sont les priors et θ_j les paramètres de nos fonctions de récompense (précédemment définis) associés à chaque *cluster*. L'idée est ensuite assez intuitive et consiste en trois étapes principales (voir [3] pour plus de détails).

1. On calcule, à chaque itération t :

$$z_{ij}^t = \prod_{(s,a)\in\tau_i} \pi_{\boldsymbol{\theta}_j^t}(s,a)\rho_j^t / Z$$

à savoir la probabilité qu'une trajectoire donnée τ_i ait été générée par une intention j ; Z est la constante de normalisation et $\pi_{\boldsymbol{\theta}_j}(s,a)$ est la politique *Softmax* déterminée par la Q-valeur sous l'hypothèse temporaire $\boldsymbol{\theta}$. Elle est donc définie comme suit :

$$\pi_{\boldsymbol{\theta}_j^t}(s,a) = \frac{e^{\beta Q_{\boldsymbol{\theta}_t}(s,a)}}{\sum\limits_{a'} e^{\beta Q_{\boldsymbol{\theta}_t}(s,a')}}$$

avec $\beta \geq 0$ la constante de Boltzman.

2. On cherche à maximiser :

$$\sum_{l=1}^{m}\sum_{i=1}^{n} \log(\rho_l^t) z_{il}^t + L(\mathcal{D}|\boldsymbol{\theta}_t)$$

à savoir la probabilité que chaque trajectoire appartienne à un cluster donné et qu'à ce cluster soit associée une fonction de récompense hypothétique. La vraisemblance (*log-likelihood*) des trajectoires sachant notre fonction de récompense est calculée comme suit :

$$L(\mathcal{D}|\boldsymbol{\theta}_t) = \sum_{l=1}^{m}\sum_{i=1}^{n} \log(\Pr(\tau_i|\theta_l^t)) z_{il}^t$$

3. On met à jour $\boldsymbol{\theta}$ par la méthode de gradient :

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t \nabla L(\mathcal{D}|\boldsymbol{\theta}_t)$$

où $\alpha_t > 0$ est, à l'instant t, le pas à appliquer lors de la descente du gradient ; et on réitère (1)-(3) jusqu'à ce que le nombre d'itérations fixé soit atteint.

Nous devrions ainsi obtenir des groupes de participants dont les distributions de probabilité sur les fonctions de récompense, c'est-à-dire les stratégies attentionnelles, se ressemblent. Nous espérons ainsi pouvoir mieux définir des stratégies de l'attention inter-modale plus efficaces que d'autres.

5 Discussion et perspectives

Nous avons présenté un protocole expérimental visant à identifier les marqueurs associés aux changements d'attention focalisée inter-modale à l'aide d'algorithmes d'IRL appliqués sur des données d'activité cérébrale. L'approche novatrice de cette proposition porte non seulement sur les métriques utilisées afin d'extraire les caractéristiques attentionnelles du signal EEG (notamment des mesures court-terme dynamiques) ; mais aussi sur l'utilisation d'algorithmes d'IRL permettant, dans le cadre d'un modèle MDP, d'apprendre la fonction de récompense associée à une politique attentionnelle donnée. Le protocole expérimental que nous avons développé nous permettra par ailleurs d'identifier une population dite "experte" (celle dont les performances sont bonnes) par la fonction de récompense sous-jacente à la politique optimale qu'on y observera.

Pour inférer et classer différentes dynamiques attentionnelles nous pensons qu'un algorithme d'IRL qui combine clustering et modèle probabiliste est le choix le plus cohérent dans le cadre de cette étude. Ce choix doit cependant être nuancé par le fait que nous devrons estimer la fonction de transition $\mathcal{T}$ entre états que nous ne connaissons pas *a priori*. Par ailleurs le modèle $\mathcal{M}$ que nous avons défini n'inclut pas l'existence d'un état final absorbant puisqu'il s'agit avant tout d'une tâche de réaction en continu. Enfin il est possible qu'un agent ne suive pas une mais plusieurs fonctions de récompense variables dans le temps. À cet égard la clusterisation d'une même trajectoire en plusieurs fonctions de récompense subordonnées et l'intégration de cycles dans un modèle bayésien non paramétrique [25] est une piste complémentaire à envisager.

Quant aux perspectives de ce travail il s'agit, dans un premier temps, de tester nos hypothèses en acquérant les données EEG des trente participants prévus. À plus long terme, la définition exacte de l'activité cérébrale associée à un comportement "expert" efficace pourrait permettre : (i) de mettre en place des interfaces cerveau-machine pour détecter des états attentionnels dégradés et lancer des contre-mesures quand, par example, l'activité d'un opérateur dévie d'une activité optimale ; (ii) aux algorithmes d'intelligence artificielle de tirer profit de l'étude des heuristiques mises en place par le cerveau pour optimiser l'allocation de ses ressources ; enfin (iii) d'améliorer la formation des apprentis en comprenant comment maximiser l'utilisation de leurs réseaux attentionnels – des techniques récentes comme celle du *neurofeedback* sont à cet égard prometteuses.

Remerciements

Ce travail est financé par ANITI - *Artificial and Natural Intelligence Toulouse Institute*, Institut 3IA, ANR-19-PI3A-0004.

Références

[1] P. Abbeel and A.Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, page 1, 2004.

[2] S. Arora and P. Doshi. A survey of inverse reinforcement learning : Challenges, methods and progress. *arXiv preprint arXiv :1806.06877*, 2018.

[3] M. Babes, V.N. Marivate, K. Subramanian, and M.L. Littman. Apprenticeship learning about multiple intentions. In *ICML*, 2011.

[4] C.L. Baker, R. Saxe, and J.B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3) :329–349, 2009.

[5] A.K.R. Bauer, S. Debener, and A.C. Nobre. Synchronisation of neural oscillations and cross-modal influences. *Trends in cognitive sciences*, 2020.

[6] D. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations, 2019.

[7] E. Bullmore and O. Sporns. Complex brain networks : graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3) :186–198, 2009.

[8] M.X. Cohen and R. Gulbinaite. Rhythmic entrainment source separation : Optimizing analyses of neural responses to rhythmic sensory stimulation. *Neuroimage*, 147 :43–56, 2017.

[9] M. Corbetta, G. Patel, and G.L. Shulman. The reorienting system of the human brain : from environment to theory of mind. *Neuron*, 58(3) :306–324, 2008.

[10] F. Dehais, H.M. Hodgetts, M. Causse, J. Behrend, G. Durantin, and S. Tremblay. Momentary lapse of control : A cognitive continuum approach to understanding and mitigating perseveration in human error. *NBR*, 100 :252–262, 2019.

[11] F. Dehais, A. Lafont, R. Roy, and S. Fairclough. A neuroergonomics approach to mental workload, engagement and human performance. *Frontiers in Neuroscience*, 14 :268, 2020.

[12] F. Dehais, I. Rida, R.N. Roy, J. Iversen, T. Mullen, and D. Callan. A pbci to predict attentional error before it happens in real flight conditions. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 4155–4160, 2019.

[13] S.M. Doesburg, A.B. Roggeveen, K. Kitajo, and L.M. Ward. Large-scale gamma band phase synchronization and selective attention. *Cerebral cortex*, 18(2) :386–396, 2007.

[14] G. Durantin, F. Dehais, N. Gonthier, C. Terzibas, and D.E. Callan. Neural signature of inattentional deafness. *Human brain mapping*, 38(11) :5440–5455, 2017.

[15] D. Fougnie, J. Cockhren, and R. Marois. A common source of attention for auditory and visual tracking. *Attention, Perception, & Psychophysics*, 80(6) :1571–1583, 2018.

[16] C.S. Herrmann. Human eeg responses to 1–100 hz flicker : resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Experimental brain research*, 137(3) :346–353, 2001.

[17] M.J. Kaminski and K.J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological cybernetics*, 65(3) :203–210, 1991.

[18] M.J. Kaminski, A. Brzezicka, J. Kaminski, and K.J. Blinowska. Coupling between brain structures during visual and auditory working memory tasks. *International journal of neural systems*, 29(3), 2019.

[19] A. Khosla, P. Khandnor, and T. Chand. A comparative analysis of signal processing and classification methods for different applications based on eeg signals. *Biocybernetics and Biomedical Engineering*, 40(2) :649–690, 2020.

[20] E. Klein, B. Piot, M. Geist, and O. Pietquin. Structured classification for inverse reinforcement learning. *JMLR*, 2012 :1–14, 2013.

[21] A. Kolobov. Planning with markov decision processes : An AI perspective. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1) :1–210, 2012.

[22] J.Z. Kolter, P. Abbeel, and A.Y. Ng. Hierarchical apprenticeship learning with application to quadruped locomotion. In *Advances in Neural Information Processing Systems*, pages 769–776, 2008.

[23] D. Lakens. Calculating and reporting effect sizes to facilitate cumulative science : a practical primer for t-tests and anovas. *Frontiers in psychology*, 4 :863, 2013.

[24] G.W. Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14 :29, 2020.

[25] B. Michini and J.P. How. Bayesian nonparametric inverse reinforcement learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 148–163, 2012.

[26] G. Neu and C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods, 2012.

[27] A.Y. Ng and S.J. Russell. Algorithms for inverse reinforcement learning. In *IMSL*, volume 1, page 2, 2000.

[28] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

[29] N.D. Ratliff, J.A. Bagnell, and M.A. Zinkevich. Maximum margin planning. In *ICML*, pages 729–736, 2006.

[30] E.E. Smith and J. Jonides. Working memory : A view from neuroimaging. *Cognitive psychology*, 33(1) :5–42, 1997.

[31] R.S. Sutton and A.G. Barto. *Reinforcement learning : An introduction*. MIT press, 2018.

[32] U. Syed and R.E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in neural information processing systems*, pages 1449–1456, 2008.

[33] J.J. Todd, D. Fougnie, and R. Marois. Visual short-term memory load suppresses temporo-parietal junction activity and induces inattentional blindness. *Psychological science*, 16(12) :965–972, 2005.

[34] N.J. Williams, I. Daly, and S.J. Nasuto. Markov model-based method to analyse time-varying networks in eeg task-related data. *Frontiers in computational neuroscience*, 12 :76, 2018.

[35] B.D. Ziebart, A.L. Maas, J.A. Bagnell, and A.K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438, 2008.